Utiliser une base distributionnelle pour filtrer un dictionnaire de synonymes

François Morlane-Hondère
CLLE-ERSS, Université de Toulouse - Le Mirail, 5, allées Antonio Machado - Toulouse Cedex 9
francois.morlane@univ-tlse2.fr

RÉSUMÉ

Cette étude vise à mettre en lumière l'intérêt qu'il peut y avoir à se servir d'une ressource générée par analyse distributionnelle automatique pour orienter les résultats fournis par un dictionnaire de synonymes. En croisant une base distributionnelle calculée à partir d'un corpus constitué d'articles de l'encyclopédie Wikipédia et le Dictionnaire Électronique des Synonymes du CRISCO, nous montrons qu'une partie seulement des synonymes proposés pour un mot donné partagent ses contextes d'apparition. Nous mettons au jour plusieurs raisons qui expliquent ce phénomène. Nous montrons ensuite que ce décalage s'observe différemment selon la nature du corpus qui a permis de calculer la base distributionnelle qui sert à filtrer le dictionnaire. Cela signifie que la nature du corpus oriente le type de synonymes filtrés par la base distributionnelle. Nous envisageons d'appliquer ce principe dans un système de réorganisation dynamique des synonymes du dictionnaire du CRISCO.

ABSTRACT

Using distributional analysis for synonym selection

In this study, we aim to highlight the benefits of using a distributional resource to improve the sorting of the synonyms contained in a dictionary. We compare a distributional resource which was created from a Wikipedia corpus and the Electronic Dictionary of Synonyms of CRISCO. We show that only a part of a given word's synonyms share its contexts in the Wikipedia corpus. We find several explanations for this phenomenon. Then, we compare the dictionary with other distributional resources and we show that the nature of the corpus affects the range of the discrepancy. That means that the nature of the corpus has an influence on the kind of synonyms that appear in the distributional resources. We plan to use this principle for the automatic reorganization of the dictionary's synonyms.

MOTS-CLÉS : analyse distributionnelle, synonymie, corpus, sémantique lexicale.

KEYWORDS: distributional analysis, synonymy, corpus, lexical semantics.

1 Introduction

Cette étude s'inscrit dans la problématique globale de l'évaluation des bases lexicales générées par analyse distributionnelle automatique. Développé dans Harris (1968), le principe de l'analyse distributionnelle (désormais *AD*) s'appuie sur l'idée selon laquelle le degré de similarité sémantique entre deux mots peut se mesurer sur la base du recouvrement de leurs contextes syntaxiques. Il a été automatisé dès le début des années 90 et a été notamment utilisé par la

suite pour assister la création d'ontologies (Grefenstette, 1994; Habert, 1998). Toutefois, les limitations propres aux outils d'analyse et le manque de données ne permettait réellement que d'entrevoir les possibilités de la méthode distributionnelle.

Couplé à l'arrivée d'analyseurs syntaxiques dits "robustes", l'accroissement de la quantité de textes accessibles au format électronique a été un facteur important dans la popularisation de la méthode distributionnelle. Ce phénomène a en effet permis de répondre à la nécessité pour l'AD de disposer de grandes quantités de données. Des ressources constituées d'archives de journaux, d'œuvres littéraires ou de textes issus du Web permettent aujourd'hui d'appliquer le principe distributionnel à grande échelle, sur des corpus de différentes natures.

Cette diversification rend encore plus prégnante la question du type de relations qui sont mises au jour ainsi que l'impact du corpus sur les bases distributionnelles générées. Différents travaux ont cherché à évaluer l'efficacité des méthodes distributionnelles pour le repérage des relations lexicales en employant des lexiques de référence (Turney, 2008; Baroni et Lenci, 2011) ou des tâches sémantiques (Baroni et Lenci, 2010). Dans des études précédentes, nous nous sommes notamment intéressés au cas de l'antonymie (Morlane-Hondère et Fabre, 2010) et de la méronymie (Morlane-Hondère et Fabre, 2012). Tous ces travaux confirment la grande diversité des relations que peut détecter l'AD, et montrent aussi la nécessité de mieux comprendre sous quelles conditions le critère distributionnel opère.

L'aspect évaluatif est destiné à prendre davantage d'importance avec la tendance actuelle à puiser des textes sur le Web (Turney, 2008; Baroni et Lenci, 2010) ou à fusionner plusieurs ressources de natures différentes (Baroni et Lenci, 2010). Cette démarche se fait au détriment d'une réflexion pourtant nécessaire autour de la caractérisation des corpus pour le calcul distributionnel. On sait en effet que les rapprochements produits par l'AD sont le reflet des fonctionnements des mots en corpus. Or, générer des bases distributionnelles à partir de corpus hétérogènes et/ou constitués de textes dont la nature est inconnue peut compliquer l'évaluation des ressources générées en gênant l'interprétation des résultats produits. Cela concourt à augmenter l'aspect *boîte noire* de la méthode, qui perd par conséquent de son intérêt du point de vue de l'analyse linguistique.

Nous proposons ici d'aborder cette double problématique de la caractérisation des résultats produits par cette méthode ainsi que celle de l'influence des corpus qu'elle prend en entrée à travers l'étude des effets du filtrage d'un dictionnaire de synonymes par une ressource distributionnelle. Notre démarche s'appuie sur l'hypothèse selon laquelle une base distributionnelle peut permettre de *sélectionner* les synonymes d'un dictionnaire en présentant à l'utilisateur ceux qui sont les plus immédiatement substituables à un mot donné (section 2). Après avoir présenté les ressources distributionnelles utilisées ainsi que le protocole que nous avons mis en œuvre – sections 3 et 4 –, nous procédons à une analyse linguistique des effets du filtrage par une base distributionnelle calculée à partir d'un corpus encyclopédique (section 5). Nous décrivons ensuite l'impact de la nature du corpus sur le filtrage des synonymes en croisant le dictionnaire et deux autres bases distributionnelles calculées à partir de corpus journalistique et littéraire (section 6).

2 Intérêts du filtrage

Le lexique de synonymes avec lequel nous travaillons est le Dictionnaire électronique des synonymes (*DES* ; Manguin *et al.*, 2004), développé au sein du CRISCO (Université de Caen). Il regroupe les synonymes relevés dans sept dictionnaires (dictionnaires analogiques et dictionnaires

de synonymes) – le *Bailly*, le *Benac*, le *Du Chazaud*, le *Guizot*, le *Lafaye*, le *Larousse* et le *Robert*. La version dont nous disposons comporte 389 182 paires de noms (52 %), verbes (26 %) et adjectifs (22 %) [1], mots simples ou syntagmes. Il est à noter que les paires ont été symétrisées (pour une relation *A/B* a été générée une relation *B/A*) : le dictionnaire compte donc 194 576 relations de synonymie réciproques. Le DES étant régulièrement mis à jour, il comptait, au 25 octobre 2012, 200 849 relations réciproques. Ce dictionnaire a été précédemment employé dans des travaux comme ceux de Bourigault et Galy (2005) ou Morlane-Hondère et Fabre (2012) pour évaluer des bases distributionnelles.

Lancée en 1998, la version en ligne du DES a connu un succès grandissant, si bien qu'en octobre 2012, le site reçoit 200 000 requêtes par jour (Orosemane, 2012). Le dictionnaire peut être consulté soit *via* le site du CRISCO [2], soit sur la plate-forme CNRTL [3]. Or, ce lexique, consultable sur le Web, constitue pour de nombreux utilisateurs un outil d'aide à la rédaction. Manguin (2005) rapporte qu'il est notamment utilisé comme tel par les administrations québécoises et suisses, qui sont parmi les plus grands utilisateurs du DES. L'article de Orosemane (2012) témoigne également de son utilisation dans le milieu journalistique. La profusion de synonymes fournis par le DES pour un mot donné est toutefois perçue, par certains utilisateurs, comme une gêne à son utilisabilité (Thot Cursus, 2012).

L'ordre dans lequel sont présentés les synonymes est défini selon un calcul de proximité basé sur le principe suivant :

> [S]i un synonyme recouvre beaucoup de sens élémentaires du mot-vedette, il est assez proche de ce dernier au point de vue sémantique. (Manguin *et al.*, 2004, p. 6)

Il s'agit donc de s'appuyer sur le réseau que forment les synonymes pour faire apparaître en début de liste ceux qui partagent le plus de cliques avec le mot vedette. Autrement dit, le site fournit en premier lieu les synonymes les plus probables.

Ce mode de présentation pose un problème majeur : l'ordre dans lequel sont présentés les synonymes est statique, dans le sens où il ne prend pas en compte le contexte dans lequel se trouve le mot dont l'utilisateur cherche un synonyme. Sans cette information, il est impossible de prédire quels synonymes répondront le mieux au besoin de l'utilisateur. Or, on sait que tous les synonymes d'un mot ne sont pas également adaptés à un contexte donné (Murphy, 2003) : par exemple, *repli* sera un meilleur synonyme de *retrait* dans un texte d'actualité qui évoque la progression d'une force armée alors que *reflux* sera plus adapté à un texte relatif au domaine maritime, de la même façon qu'*abolition* sera plus pertinent dans un texte de loi ou *amputation* dans un texte relatif à la médecine. Le fait de mettre en avant les synonymes les plus probables constitue donc un pis-aller face à l'absence d'informations sur le contexte d'apparition du mot qui a été donné en requête (ou du contexte dans lequel va s'insérer le synonyme).

L'AD permet de synthétiser les contextes d'apparition d'un mot dans un corpus donné et de rapprocher les mots qui partagent ces contextes. Nous proposons donc ici d'utiliser cette méthode pour filtrer dynamiquement les synonymes proposés par le DES. Ainsi, plutôt que de renvoyer à l'utilisateur les synonymes les plus probables du mot vedette étant donné un réseau de synonymes construit *in abstracto*, il s'agit de renvoyer les synonymes les plus pertinents pour un type de corpus donné. À titre d'illustration, nous avons rapporté ci-dessous les synonymes du nom

1. Ces proportions s'appuient sur un étiquetage automatique accompagné d'une vérification manuelle effectués par Mai Ho-Dac et Franck Sajous (CLLE-ERSS).
2. `http://www.crisco.unicaen.fr/des/`
3. `http//www.cnrtl.fr/synonymie/`

commission. Les mots qui ont été rapprochés de *commission* suite à l'analyse d'un corpus constitué d'articles de Wikipédia apparaissent en gras :

> *prime, courtage, rémunération, gain, salaire, attribution, remise,* ***intérêt, délégation, comité, charge, traitement, réunion, pouvoir****, paiement,* ***mission, mandat, message, titre, groupe, course****, change,* ***bureau, brevet, besoin, tribunal***

On voit ainsi que les mots *prime*, *rémunération*, *gain*, *salaire* et *paiement*, bien qu'ils apparaissent dans le corpus, ne partagent pas suffisamment de contextes d'apparition avec *commission* pour être captés comme des voisins distributionnels. Or, ces mots relèvent d'une acception de *commission* bien particulière, celle de "Rétribution perçue par le commissionnaire" [4]. On peut donc en déduire que dans un corpus constitué d'articles de Wikipédia, le mot *commission* n'est pas – ou est peu – employé dans des contextes où il renvoie à une valeur pécuniaire. En revanche, il partage les mêmes contextes que :

– *délégation*, *comité*, *groupe* ou *bureau*, qui renvoient à l'acception "Ensemble de personnes officiellement chargées d'une mission à caractère public",
– *charge*, *message*, *mission* ou *mandat*, qui relèvent du sens "Charge qu'une personne reçoit de faire quelque chose".

On se base donc ici sur une AD pour extraire les synonymes qui renvoient au mot vedette tel qu'il est employé dans le corpus.

3 Les voisins distributionnels

Nous appelons *voisins distributionnels* les ressources distributionnelles générées au sein du laboratoire CLLE-ERSS par la chaîne de traitement Syntex-Upéry.

Développé par Didier Bourigault (2007), Syntex est un analyseur syntaxique en dépendance. Les analyses qu'il produit servent d'entrée à Upéry (Bourigault, 2002), qui est le module qui procède au calcul des voisins. Elles sont dans un premier temps ramenées sous la forme de triplets syntaxiques <*mot1*, RELATION, *mot2*>. Par exemple, on obtient les triplets suivants à partir de l'analyse de la phrase *Les paysans révolutionnaires détruisirent le couvent* (seuls les triplets dont *mot1* et *mot2* sont des noms, des verbes, des adjectifs ou des syntagmes nominaux sont retenus) :

– *détruire*, SUJ, *paysan*
– *détruire*, OBJ, *couvent*
– *paysan*, MOD, *révolutionnaire*

Pour les besoins du calcul distributionnel, les triplets sont d'abord réduits sous la forme de couples de type <*mot1*, *mot2*_REL>. Dans la terminologie d'Upéry, on parle de prédicats et d'arguments :

– le prédicat résulte de la concaténation du recteur et de l'étiquette de la relation,
– l'argument correspond au mot qui est régi.

Les résultats de la conversion des exemples précédents au format <*argument*, *prédicat*> sont donc les suivants :

– <*paysan*, *détruire*_SUJ>
– <*couvent*, *détruire*_OBJ>
– <*révolutionnaire*, *paysan*_MOD>

On obtient ainsi deux types d'entités : des prédicats dont les contextes sont tous les arguments

4. Les définitions que nous donnons sont issues du TLFi (`http://atilf.atilf.fr/`).

avec lesquels ils apparaissent dans le corpus, et *vice versa*. Ce formalisme va donner lieu à des rapprochements distributionnels qui opèrent à deux niveaux :
– les arguments sont rapprochés entre eux en fonction des prédicats qu'ils partagent.

couple d'arguments	prédicats partagés
paysan/ouvrier	*travailler*_SUJ, *recruter*_OBJ...
couvent/temple	*bâtir*_OBJ, *piller*_OBJ...

– réciproquement, les prédicats sont rapprochés entre eux en fonction des arguments partagés.

couple de prédicats	arguments partagés
*détruire*_OBJ/*construire*_OBJ	*immeuble*, *bâtiment*...

Étant donné le filtrage sur les catégories grammaticales qui a été opéré en amont, on peut recenser trois types de prédicats : les prédicats verbaux, nominaux et adjectivaux. Les relations qu'ils peuvent porter sont les relations sujet ou objet (prédicats verbaux), modifieur (prédicats adjectivaux) ou *prep* (tous types de prédicats), qui apparaît quand une expansion prépositionnelle est rattachée au nom, au verbe, ou à l'adjectif. Dans ce cas, c'est la préposition qui est accolée au recteur dans le prédicat (<*canon*, *détruire*_AVEC>).

L'étape suivante consiste à associer à chaque prédicat un vecteur constitué de l'ensemble des arguments qu'il prend dans le corpus (et réciproquement). Ces vecteurs sont ensuite comparés les uns aux autres afin de rapprocher les prédicats/arguments qui ont la similarité distributionnelle la plus élevée. La mesure de similarité qui est utilisée dans les voisins distributionnels est l'indice de Lin (1998). Le score de similarité de deux prédicats ou arguments varie – de 0 à 1 – en fonction de plusieurs facteurs : le nombre de contextes partagés, le nombre de triplets différents dans lesquels chacun de deux mots apparaît (indice de productivité), le degré de spécificité du contexte qui permet d'effectuer le rapprochement (se reporter à Bourigault (2002) pour les détails de la procédure de calcul).

Dans le cadre de cette étude, nous utilisons trois ressources distributionnelles qui ont été produites par la chaîne Syntex-Upéry à partir de trois corpus de différentes natures :
– un corpus d'environ 262 millions de mots constitué de l'intégralité des articles de l'encyclopédie en ligne Wikipédia [5] dans sa version de juin 2008,
– un corpus comprenant l'ensemble des articles parus dans le journal Le Monde sur une période de 10 ans – de 1991 à 2000 – soit environ 200 millions de mots,
– un corpus d'environ 30 millions de mots constitué de 515 romans datant du XX^{e} siècle issus de la base Frantext [6].

Ces trois corpus ont donc permis de générer les ressources appelées *voisins de Wikipédia* (désormais *VDW*), *voisins de Le Monde* (*VDLM*) et *voisins de Frantext* (*VDF*).

La taille de ces corpus paraît modeste en comparaison de ceux qui ont pu être utilisés dans des travaux comme ceux de Turney (2008) ou Baroni et Lenci (2010) qui utilisent les corpus issus du Web, dont le nombre de mots se compte en milliards. La raison pour laquelle nous avons malgré tout travaillé avec ces ressources est double. D'une part, elles sont les seules actuellement disponibles au laboratoire CLLE-ERSS (qui est en train de se doter de bases distributionnelles

5. http://fr.wikipedia.org/
6. http://www.frantext.fr

calculées à l'aide de l'analyseur *open source* Talismane [7] (Urieli et Tanguy, 2013)). D'autre part, nous nous situons dans une démarche qui s'appuie notamment sur des retours au contexte pour expliquer les phénomènes observés. Or, comme nous l'avons évoqué dans l'introduction, l'utilisation de corpus issus du Web complique ce type d'approche. Les corpus que nous utilisons ont donc l'avantage qu'ils appartiennent à des genres identifiés et connus.

La différence de taille entre les trois corpus utilisés se répercute sur celle des trois bases de voisins. Sont considérés comme des *voisins* tous les couples de prédicats ou d'arguments dont l'indice de Lin est supérieur ou égal à 0,1. On compte ainsi :
– 3 922 657 couples dans les VDW,
– 5 525 480 couples dans les VDLM,
– 792 356 couples dans les VDF.

Ces trois bases sont consultables sur la plate-forme REDAC [8].

4 Sélection des données à analyser

Notre approche consiste à sélectionner des mots vedettes du DES et à analyser, parmi tous les synonymes auxquels ils sont associés dans le lexique, ceux qui sont – ou ne sont pas – captés comme des voisins distributionnels dans les VDW, VDLM et VDF. Il nous sera ainsi possible d'étudier la façon dont les corpus orientent le repérage de certains synonymes.

Si l'on mesure successivement le recouvrement entre ces trois bases de voisins et le DES en ne prenant en compte que le lexique qui est partagé par les ressources comparées, la proportion de synonymes dans les VDW, VDLM et VDF est respectivement de 1,8 %, 1,5 % et 3,5 %. Réciproquement, la proportion de synonymes du DES reconnus comme des voisins distributionnels est de 41,6 % quand la comparaison se fait avec les VDW, de 33,7 % avec les VDLM et de 21,1 % avec les VDF. Ainsi, seulement un tiers des couples de synonymes du DES sont analysés comme des voisins distributionnels dans nos ressources. Cela laisse supposer que certains couples de synonymes ne sont pas suffisamment substituables en corpus pour que l'AD puisse les rapprocher.

Une des raisons à cette non-substituabilité peut être liée à des propriétés statistiques des couples. En effet, Weeds (2003) montre que les mesures de similarité utilisées pour comparer les vecteurs de contextes ont tendance à rapprocher davantage les mots qui ont des fréquences comparables. De ce fait, une certaine proportion des couples de synonymes – ceux qui sont constitués de mots dont les fréquences sont déséquilibrées – ont statistiquement moins de chance d'être captés par l'AD. Par la suite, nous avons cherché à écarter ces couples de nos échantillons : les couples de synonymes que nous avons voulu étudier en priorité sont ceux qui n'étaient pas sujets au biais évoqué par Weeds. Autrement dit, ceux qui avaient toutes les chances d'être identifiés par l'AD mais qui ne l'ont pas été.

Nous avons distingué les couples les plus susceptibles d'être extraits par l'AD de ceux qui sont pénalisés en calculant le rapport entre la productivité – *r_prod* – de leurs deux membres. La productivité, que nous définissons comme le nombre de contextes différents dans lesquels apparaît un mot, nous semble en effet plus représentative de la distribution d'un mot que sa fréquence (ces deux valeurs sont toutefois très liées (Morlane-Hondère et Fabre, 2012)). Par

7. Il est librement téléchargeable à l'adresse suivante : `http://redac.univ-tlse2.fr/applications/talismane.html`
8. `http://redac.univ-tlse2.fr/index.html`

exemple les synonymes *monarque* et *chef* ont des productivités respectives de 59 et de 1941, leur r_prod est donc de 0,03 (soit le résultat de la division de 59 par 1941). Nous nous intéressons davantage à l'analyse de couples dont les productivités sont moins inégales comme *aspect* et *air*, qui apparaissent respectivement dans 883 et 935 contextes différents, et dont le r_prod est de 0,94.

Nous avons choisi de ne conserver que les 52 092 couples de synonymes dont le r_prod est supérieur ou égal à 0,23. La raison en est que, sur le 112 863 couples du DES dont le lexique est commun à celui des VDW, ceux qui ont un r_prod inférieur à 0,23 sont majoritairement des non-voisins. Au delà, le rapport s'inverse. Cela confirme que les couples composés de deux mots dont les productivités sont inégales (et donc dont le r_prod est faible), ont moins de chance d'être captés par l'AD.

5 Filtrage des synonymes par les VDW

Avec l'exemple de *commission*, commenté à la section 2, nous avons vu qu'un mot vedette n'était pas forcément voisin avec tous ses synonymes. Dans cette première étude, nous isolons les mots vedettes dont les synonymes sont le moins bien repérés afin d'identifier les conditions par lesquelles se fait le filtrage.

Nous avons choisi de nous intéresser aux cas les plus emblématiques du décalage entre le DES et les VDW, à savoir les mots vedettes dont le nombre de synonymes extraits est quasi nul (mais dont le nombre de synonymes total est supérieur ou égal à 10). Nous avons donc calculé la proportion de synonymes captés par mot vedette et extrait les 30 noms, les 30 verbes et les 30 adjectifs dont la proportion est la plus basse, c'est-à-dire inférieure à 0,1 %. Ces derniers constituent notre échantillon d'étude. Nous en avons rapporté un extrait ci-dessous (le chiffre entre parenthèses correspond au nombre de synonymes qui n'ont pas été extraits comme des voisins, soit la quasi totalité d'entre eux) :

- noms : *éclat* (29), *remède* (21), *espérance* (20), *flamme* (20), *pli* (18), *vide* (17), *ardeur* (17), *clôture* (14), *achèvement* (14), *attente* (14)
- verbes : *exciter* (29), *vider* (22), *coller* (22), *ôter* (20), *arracher* (20), *ébranler* (17), *calmer* (17), *altérer* (17), *allonger* (16), *revêtir* (15)
- adjectifs : *fou* (30), *doux* (27), *rude* (27), *ferme* (25), *ardent* (19), *vague* (18), *habile* (16), *brut* (16), *honnête* (16), *épais* (15)

Cette section est consacrée à la description des trois causes de décalage entre le DES et les VDW que l'analyse de notre échantillon nous a permis d'identifier, à savoir la polysémie (5.1), la connotation (5.2) et la dénotation périphérique (5.3). Les exemples que nous analysons sont ceux qui, dans notre échantillon, nous ont paru les plus représentatifs des phénomènes mis au jour.

5.1 La polysémie

Le phénomène de polysémie correspond à l'explication la plus simple du non-repérage de certains synonymes. En effet, quand deux mots sont synonymes, c'est toujours en fonction d'une acception donnée (Lehmann et Martin-Berthet, 2011). Si cette acception ne se manifeste pas – ou peu – dans le corpus, alors les deux mots ne partageront pas suffisamment de contextes pour être extraits par l'AD. Cette situation peut s'observer sur de nombreux mots de notre échantillon

comme *pli*, *clôture* ou *fraternité* :
– *pli* s'emploie comme objet des verbes *oblitérer*, *affranchir* ou *poster*. Le sens qui émerge est celui de "enveloppe renfermant une lettre", ses premiers voisins sont *levée*, *courrier*, *paquet*. Ses synonymes *arête*, *repli*, *sillon*, *etc.* ont donc peu de chance d'être captés.
– *clôture* s'emploie principalement en expansion de noms d'évènements comme *jour*, *gala* ou *cérémonie*. Ses contextes d'apparition sont incompatibles avec des synonymes comme *muraille*, *grille* ou *barrage*.
– *fraternité* émerge dans le sens de "communauté ou groupement, laïc ou religieux" dans des contextes comme *rejoindre*_OBJ, *membre*_DE ou encore *fondateur*_DE alors que ses synonymes – *charité*, *sympathie*, *confiance*, *etc.* – réfèrent pour la plupart à son acception "sentiment de solidarité et d'amitié".

L'influence du corpus s'observe de façon plus systématique lors de l'étude des mots *complication* ou *anormal*, qui nous a permis d'identifier un phénomène récurrent lié à la présence de textes relevant du domaine médical dans le corpus Wikipédia. En effet, dans le corpus Wikipédia, ces mots sont utilisés comme des termes médicaux. Ils apparaissent donc dans des contextes relevant du domaine de la médecine, ce qui bloque le repérage de leurs synonymes, qui réfèrent à des acceptions non spécialisées. Par exemple, *complication* apparaît dans des contextes comme *dépister*_OBJ, *mourir*_DE, *prévenir*_OBJ et est modifié par *neurologique*, *cardio-vasculaire* ou *hépatique*. Cela montre clairement que *complication* adopte ici le sens de "phénomènes pathologiques nouveaux résultant de l'évolution d'une maladie et appelant généralement un traitement particulier". On voit donc aisément pourquoi les synonymes que renvoie le DES pour *complication*, à savoir *piège*, *chaos*, *labyrinthe* ou *combinaison* ne sont pas captés comme des voisins : les contextes dans lesquels *complication* apparaît sont tellement spécifiques que sa substitution par l'un de ses synonymes est inenvisageable. On peut également citer le cas de l'adjectif *anormal*, qui modifie principalement des noms comme *hémoglobine*, *saignement*, *gène* ou *protéine*, qui sont des contextes incompatibles avec des synonymes comme *bizarre*, *paradoxal* ou *insolite*. Ainsi, le fait que le corpus Wikipédia contienne une certaine proportion de textes relevant du domaine médical a une influence sur le sens de ces mots, qui prennent alors un sens spécialisé. En conséquence, leur distribution est radicalement différente de celles de leurs synonymes dans le DES.

Nous avons également considéré comme un phénomène de polysémie le cas des emplois métaphoriques. Ce cas de figure s'observe sur un mot vedette comme *flamme*. Le nom *flamme* apparaît en effet dans des contextes comme *lécher*_SUJ, *brûler*_AVEC ou *cerner*_SUJ, c'est-à-dire dans son acception "Mélange gazeux en combustion, dégageant de la chaleur et généralement de la lumière, produit par une matière qui brûle". Ici, le décalage naît du fait que la plupart de ses synonymes – *passion*, *désir*, *enthousiasme*, *etc.* – renvoient à un sens métaphorique de *flamme*. Ces derniers partagent dans le corpus Wikipédia des contextes comme *manifester*_OBJ, *susciter*_OBJ ou *provoquer*_OBJ, dans lesquels *flamme* n'apparaît pas.

5.2 La connotation

Nous avons observé dans notre échantillon deux types de connotations que Kerbrat-Orecchioni (1977) désigne sous le nom de "connotation énonciative" et "connotation stylistique".

Dans le premier cas, le décalage naît du fait que le mot vedette ou son synonyme porte une valeur axiologique que l'autre n'a pas. Ce phénomène est dit "énonciatif" dans le sens où l'emploi d'un mot axiologiquement marqué plutôt que d'un mot *neutre* donne des indications "non sur le référent du message, mais sur son énonciateur" (Kerbrat-Orecchioni, 1977, p. 104). Les mots

marqués et non marqués partagent le même sens dénotatif, ils peuvent donc tout de même constituer des synonymes potentiels (Murphy, 2003). Parmi les mots vedettes de notre échantillon, les adjectifs *sommaire* et *fou* nous permettent d'illustrer ce phénomène. Ils ont respectivement pour synonymes *simpliste* et *stupide*, lesquels sont clairement péjoratifs. Ils n'ont pas été extraits par l'AD. Si l'on regarde de plus près le cas de *fou*/*stupide*, on peut voir que les premiers voisins de *fou* sont *malade*, *immortel* et *endormi* et que ceux de *stupide* sont *intelligent* et *curieux*. Cela confirme le fait que *fou* est utilisé comme un adjectif non marqué alors que *stupide* est rapproché d'adjectifs qui portent une valeur subjective. Il est à noter que, puisque la rédaction des articles de Wikipédia est soumise au respect de la neutralité de point de vue [9], qui proscrit l'emploi de mots connotés, un mot comme *stupide* apparaît majoritairement dans des citations ou des titres d'œuvres. De ce fait, la spécificité de ces contextes peut créer un décalage distributionnel entre les synonymes axiologiquement marqués et leurs équivalents non marqués. Ces cas de décalage sont ainsi révélateurs d'un certain degré d'hétérogénéité du corpus Wikipédia.

Le deuxième type de connotation – la connotation stylistique – renvoie au registre de langue auquel appartient un mot. Nos données montrent que si deux mots qui appartiennent à des registres différents peuvent présenter des "propriétés sémiotiques" identiques, cela n'implique pas qu'ils seront substituables dans tous les contextes. En effet, certains mots de la langue *standard* peuvent prendre des emplois argotiques, comme c'est le cas de *veine*, qui figure dans le DES comme un synonyme de *hasard*. Or, *veine* s'emploie dans le corpus dans des contextes comme *couler*_DANS, *injection*_DANS, *occlusion*_DE ou peut être modifié par les adjectifs *splénique*, *fémoral*, *jugulaire*, *etc.* Ces contextes nous montrent clairement que *veine* est employé dans le corpus comme un type de vaisseau sanguin et non au sens de "Chance, fortune". Il n'apparaît pas dans les même contextes que *hasard* comme *part*_DE, *jeu*_DE, *rencontrer*_PAR, *etc.* On peut tirer les mêmes conclusions pour le mot vedette *chauffer*, qui, dans le DES, a pour synonymes *voler* et *dérober*, qui renvoient à une acception argotique du mot.

5.3 La dénotation périphérique

À la suite de Murphy (2003), nous distinguons dans le sens dénotatif d'un mot son noyau sémantique (*core meaning*) et ses – éventuels – traits périphériques (*peripheral features*). La notion de trait périphérique nous permet en effet d'aborder le problème des synonymes qui partagent un même noyau de sens (ce qui n'était pas le cas dans la section 5.1 : un *pli* (postal) n'est pas la même chose qu'un *repli*) mais qui se distinguent du point de vue des nuances sémantiques qu'ils peuvent véhiculer (qui sont donc exprimées par ces traits).

Parmi les mots vedettes de notre échantillon, le verbe *vider* est représentatif de ce phénomène. Les synonymes qui sont listés par le DES pour *vider* apportent des précisions sur l'action exprimée par le verbe :
– *épuiser* : "Vider (quelque chose) de son contenu ou de sa substance",
– *écoper* : "Vider l'eau qui s'accumule au fond d'une embarcation non pontée, à l'aide d'une écope",
– *évacuer* : "Vider (un pays, un lieu) des personnes qui l'occupent",
– *déménager* : "Vider (le meuble, la pièce) de tout ce qu'il contient",
– *assécher* : "Vider de ses ressources".

Les traits véhiculés par ces synonymes apportent des précisions soit sur la nature des compléments

9. http://fr.wikipedia.org/wiki/Wikipédia:Neutralité_de_point_de_vue

*vider*_OBJ	*épuiser*_OBJ	*évacuer*_OBJ	*déménager*_OBJ	*assécher*_OBJ
chargeur	deux édition	ville	siège social	marécage
poubelle	pioche	eau	franchise	marais
cuve	gisement	habitant	local	étang
récipient	stock	chaleur	colonel	lac
caisse	munition	population	fois	rivière
coffre	réserve	personne	rue	air
cache	recours	partie	siège	zone
sac	carburant	troupe	capitale	terre
querelle	ressource	territoire	mois	partie
cave	combustible	île	centre	

TABLE 1 – Différences dans la distribution de *vider* et de quelques-uns de ses synonymes.

du verbe, soit sur la manière dont s'effectue l'action. On peut supposer que ces nuances se répercutent sur la distribution de ces synonymes. Ainsi, afin d'observer ce phénomène, nous avons rapporté au tableau 1 les dix arguments [10] qui ont l'information mutuelle la plus élevée avec *vider* et ses synonymes *épuiser*, *évacuer*, *déménager* et *assécher* lorsqu'ils portent la fonction OBJ [11].

Les contextes recensés dans le tableau 1 confirment l'hypothèse que nous avons formulée plus haut, à savoir que les traits portés par les synonymes de *vider* ont une influence sur leurs distributions dans le corpus. Ainsi, alors que *vider* prend pour objets des noms de contenants (*récipient*, *caisse*, *coffre*), *épuiser* s'emploie avec des noms de ressources (*carburant*, *combustible*... *ressource*) ou de *lieux* contenant ces ressources (*stock*, *réserve*, *gisement*), *évacuer* avec des noms de lieux (*ville*, *territoire*, *île*) ou d'animés (*habitant*, *population*, *personne*), *déménager* avec des noms de lieux (*local*, *rue*, *capitale*) et *assécher* avec des noms de lieux qui contiennent habituellement de l'eau (*marécage*, *étang*, *lac*). De ce fait, alors que *vider* et ses synonymes partagent un même noyau de sens, on s'aperçoit que le chevauchement entre les noms qu'ils prennent comme objets reste très faible. Ils semblent donc davantage fonctionner sur le mode de la complémentarité plutôt que sur celui de la substituabilité : *évacuer*, *épuiser*, *déménager* et *assécher* portent des traits périphériques qui font que ces mots s'emploient dans des contextes spécifiques. Le fait que *vider* ne porte pas de trait périphérique semble bloquer son apparition dans ces contextes, donc sa substituabilité avec ses synonymes.

Les trois cas de figure que nous avons commentés dans cette section 5 nous ont permis de mettre en lumière certains des fonctionnements qui entraînent l'absence d'un synonyme dans une base de voisins distributionnels. En excluant les synonymes qui ne sont pas – ou très peu – substituables avec un mot vedette donné, l'AD permet de faire émerger ceux qui sont le plus susceptibles d'apparaître dans le contexte du mot vedette, autrement dit, les plus pertinents pour un utilisateur du DES. De ce fait, une hypothétique intégration du DES aux pages de modification des articles de Wikipédia gagnerait à proposer en priorité à l'utilisateur les synonymes sélectionnés par les VDW (au détriment des synonymes qui, par exemple, renvoient à des acceptions du mot vedette qui ne se manifestent pas dans le corpus).

10. Sauf pour *assécher*_OBJ, qui n'en compte que 9 au total.

11. Nous n'avons pas fait figurer les contextes du verbe *écoper* étant donné qu'il n'apparaît dans le corpus que dans le sens de "Subir (des dommages matériels), recevoir (des coups) ; être atteint ou touché (par quelque chose que l'on subit)".

6 Variation du repérage en fonction du corpus

Nous avons décrit, dans la section précédente, quelques unes des propriétés qui régissent la substituabilité des mots vedettes et de leurs synonymes. Dans cette section, nous montrons que la nature des synonymes qui sont rapprochés du mot vedette par l'AD varie selon que la base distributionnelle utilisée ait été calculée à partir de textes encyclopédiques, journalistiques, ou littéraires. Nous avons donc, dans un premier temps, comparé successivement la proportion de synonymes repérés comme des voisins par les VDW avec celle des VDLM et des VDF.

	Synonymes voisins	Synonymes voisins partagés avec les VDW
VDLM	11,4	10
VDF	2,5	2,1

TABLE 2 – Nombres moyens de synonymes repérés dans les VDLM et les VDF.

Les résultats, rapportés au tableau 2, montrent qu'en moyenne, les VDLM extraient 11,4 synonymes par mot vedette (sur 16), ce qui correspond exactement au nombre de synonymes captés par les VDW (la différence dans la proportion de synonymes extraits évoquée à la section 4 ne s'observe donc pas sur les données seuillées). En revanche, les VDF n'en extraient que 2,5. Nous attribuons la faiblesse de ce chiffre à la taille du corpus Frantext. On voit également que la plupart des synonymes captés par les VDLM/VDF l'ont également été par les VDW. Cela montre qu'en moyenne, la plupart des synonymes qui sont extraits par les VDLM et les VDF le sont aussi par les VDW. On note que la réciproque n'est pas vraie pour les VDF puisque les VDW reconnaissent en moyenne 4,6 fois plus de synonymes comme des voisins distributionnels que les VDF. Le fait qu'on observe un recouvrement important entre les VDW et les VDLM montre que les fonctionnements des mots vedettes et de leurs synonymes dans les corpus Wikipédia et Le Monde ne sont pas radicalement différents. Toutefois, ces résultats globaux ne doivent pas masquer le fait que les effets du corpus se font malgré tout ressentir de façon importante pour une certaine proportion de mots vedettes. À titre d'exemple, 16 % des 1202 mots vedettes observés possèdent au moins 5 synonymes qui n'ont été captés que par les VDW ou les VDLM. C'est notamment le cas du mot vedette *ton*, que nous décrivons ci-dessous.

Nous avons rapporté dans le tableau 3 la liste des synonymes du nom *ton* selon qu'ils sont captés par les bases de voisins (✓), qu'ils sont présents dans leur lexique mais non repérés comme des voisins (✗) ou absents de leur lexique (∅). Ce tableau appelle plusieurs observations. Tout d'abord, on constate que, sur 34 synonymes, beaucoup sont absents des lexiques des trois ressources. Cela est dû à la rigueur du filtrage opéré en amont. Ainsi, le mot *main*, qui apparaît dans la version non filtrée des VDW, est absent de la version seuillée (avec un r_prod à 0,23), qui ne contient que les couples les plus susceptibles d'être captés. Il est à noter que nous n'avons pas fait apparaître les synonymes qui n'apparaissent dans aucun des lexiques des trois ressources.

Le troisième point est le plus important pour notre problématique. En effet, on remarque que même si les listes de synonymes captés par les VDW et les VDLM sont de tailles à peu près identiques, leur contenu est relativement différent :

– moins de la moitié des synonymes reconnus respectivement par les VDW et les VDLM sont communs aux deux ressources. Ces synonymes sont les suivants : *accent*, *écriture*, *goût*, *manière*, *son*. Ces mots partagent donc les mêmes contextes d'apparition que *ton* que ce soit dans le

synonyme	VDW	VDLM	VDF	synonyme	VDW	VDLM	VDF
accent	✓	✓	✓	note	✓	✗	✗
air		✓	✓	nuance	✓	✗	
bruit	✗	✗	✗	parole	✓	✗	✗
corde	✗		✗	patte	✓		✗
couleur		✓	✗	plume	✓	✗	✗
écho	✗	✓		procédé	✗	✗	
écriture	✓	✓		signature	✓	✗	
expression		✓	✓	son	✓	✓	✗
façon	✓	✓	✗	style		✓	
facture	✗	✗		teinte	✓		
genre		✓	✗	tension	✗	✗	
goût	✓	✓	✗	timbre	✗	✗	
griffe	✗			tonalité	✓		
main		✗	✗	touche	✗	✗	
manière	✓	✓	✗	tour		✗	✗
mode		✗		verbe	✗		
musique		✗	✗	voix		✓	✓

TABLE 3 – Repérage des synonymes du nom *ton* en fonction des bases de voisins.

corpus Wikipédia ou dans le corpus Le Monde.

– a l'inverse, les mots *nuance*, *signature*, *parole*, *plume* et *note* ne partagent les mêmes contextes que *ton* que dans le corpus Wikipédia. On peut supposer que ces rapprochements sont à attribuer à une plus grande présence du vocabulaire des arts dans les VDW. Une comparaison des modifieurs les plus fréquents de *ton* dans les corpus Wikipédia et Le Monde va dans le sens de cette intuition : alors que *ton* est fréquemment modifié par des adjectifs comme *majeur*, *mineur*, *chaud*, *bleu* ou *clair* dans le corpus Wikipédia, on trouve parmi ses modifieurs privilégiés dans le corpus Le Monde des adjectifs comme *grave*, *ferme*, *vif*, *modéré* ou *solennel*. Dans le deuxième cas, l'emploi de *ton* privilégié semble donc être le suivant : "Qualité de la voix (hauteur, timbre, intensité)".

On a ici pu déduire de l'analyse des synonymes de *ton* captés par les différentes bases de voisins que le sens dans lequel était employé ce mot différait en fonction de la base distributionnelle considérée. Cette observation traduit une différence d'usage dans les corpus qui ont permis de générer ces ressources distributionnelles.

À ce stade, il est également possible de mener une comparaison des synonymes captés par les différentes bases de voisins qui va au delà de la dichotomie absence/présence. En effet, même si dans la plupart des cas, les synonymes extraits par les VDW et les VDLM pour un mot vedette donné sont à peu près les mêmes, il n'est pas évident que ces synonymes aient un score de similarité distributionnelle avec le mot vedette identique dans les deux ressources. À titre d'exemple, les 23 synonymes du mot vedette *tour* ont tous été captés par les VDW et les VDLM. Toutefois, on peut voir ci-dessous que les cinq meilleurs voisins de *tour* portent sur deux acceptions différentes en fonction de la ressource :

– VDW : *bâtiment* (0,47), *construction* (0,345), *pièce* (0,257), *course* (0,186), *forme* (0,166)
– VDLM : *course* (0,217), *tournée* (0,197), *façon* (0,190), *voyage* (0,170), *marche* (0,169)

De ce fait, il devient possible d'observer les effets du corpus à un niveau de finesse plus élevé que

celui que nous avons adopté plus haut lors de l'étude du repérage des synonymes de *ton*.

7 Conclusion

Dans cette étude, nous avons abordé la question de ce que pourrait apporter à un dictionnaire de synonymes un filtrage par une base distributionnelle. Cela nous a amené à montrer, dans un premier temps, que tous les synonymes d'un mot vedette donné ne partagent pas ses contextes d'apparition. Des phénomènes comme la polysémie, la connotation ou des incompatibilités en termes de traits dénotatifs font que la distribution du mot vedette et de certains de ses synonymes divergent. De ce fait, l'AD permet d'opérer une sélection parmi les synonymes proposés pour un mot vedette donné.

Nous avons ensuite montré que ce phénomène pouvait varier selon que l'on filtrait le dictionnaire avec une ressource calculée à partir d'un corpus constitué de textes encyclopédiques, journalistiques ou littéraires. La pertinence des synonymes d'un mot vedette donné n'est en effet pas absolue mais varie en fonction du type de corpus dans lequel le mot est employé.

Nous proposons par la suite d'exploiter les informations que fournit le corpus sur la distribution du mot dont on recherche le synonyme. Il s'agirait alors de se servir de ressources distributionnelles en support d'un dictionnaire de synonymes pour savoir quels sont ceux qui sont les plus pertinents pour un type de texte donné et de réorganiser les résultats fournis en conséquence. Une étape d'évaluation serait nécessaire pour mesurer les apports d'une telle démarche. Nous avons donc prévu de poursuivre la présente étude en mettant en place un protocole d'évaluation dans lequel il sera demandé à des utilisateurs de sélectionner les meilleurs synonymes pour un mot donné dans un ensemble de phrases extraites des corpus Wikipédia, Le Monde et Frantext. Nous faisons ainsi l'hypothèse que le choix des utilisateurs se portera en priorité sur les synonymes qui sont voisins du mot cible dans la base distributionnelle calculée à partir du corpus d'où a été extraite la phrase.

Références

Baroni, M. et Lenci, A. (2010). Distributional Memory : A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M. et Lenci, A. (2011). How we BLESSed distributional semantic evaluation. *In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS'11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle* (24–27 juin 2002), Nancy.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel :* Syntex. Habilitation à diriger des recherches. Université Toulouse II – Le Mirail.

Bourigault, D. et Galy, E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In 4e journées de la linguistique de corpus* (15–17 septembre 2005), Lorient.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

HABERT, B. (1998). *Des mots complexes possibles aux mots complexes existants : l'apport des corpus*. Habilitation à diriger des recherches en linguistique, Université Lille III – Charles de Gaulle.

HARRIS, Z. (1968). Mathematical structures of language. John Wiley & Sons.

KERBRAT-ORECCHIONI, C. (1977). *La connotation*. Linguistique et sémiologie. Presses universitaires de Lyon.

LEHMANN, A. et MARTIN-BERTHET, F. (2011). *Introduction à la lexicologie : sémantique et morphologie*. Collection Lettres supérieures. Armand Colin.

LIN, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

MANGUIN, J.-L. (2005). Les dictionnaires en ligne : nouvelles diffusions, nouveaux objectifs. Paris, France.

MANGUIN, J. L., FRANÇOIS, J., EUFE, R., FESENMEIER, L., OZOUF, C. et SÈNÈCHAL, M. (2004). Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux. *In Cahiers du CRISCO*, volume 34. CRISCO, Université de Caen.

MORLANE-HONDÈRE, F. et FABRE, C. (2010). L'antonymie observée avec des méthodes de TAL : une relation à la fois syntagmatique et paradigmatique ? *In* ASSOCIATION POUR LE TRAITEMENT AUTOMATIQUE DES LANGUES (ATALA), éditeur : *Actes de TALN 2010*, page 6, Montréal, Canada. Article court.

MORLANE-HONDÈRE, F. et FABRE, C. (2012). Étude des manifestations de la relation de méronymie dans une ressource distributionnelle (Study of Meronymy in a Distribution-Based Lexical Resource) [in French]. *In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 169–182, Grenoble, France. ATALA/AFCP.

MORLANE-HONDÈRE, F. et FABRE, C. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. *SHS Web of Conferences*, 1:1001–1015.

MURPHY, M. L. (2003). *Semantic Relations and the Lexicon : Antonymy, Synonymy and other Paradigms*. University Press, Cambridge.

OROSEMANE, L. (2012). Qui se cache derrière le dictionnaire des synonymes de Caen ? *Rue 89*. En ligne. `http://www.rue89.com/2012/10/28/qui-se-cache-derriere-le-dictionnaire-des-synonymes-de-caen-236552` Page consultée le 25/02/2013.

THOT CURSUS (2012). Un dictionnaire des synonymes gros comme ça ! En ligne. `http://cursus.edu/dossiers-articles/articles/8902/dictionnaire-des-synonymes-gros-comme/` Page consultée le 29/03/2013.

TURNEY, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. *In COLING*, pages 905–912.

URIELI, A. et TANGUY, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane (à paraître). *In Actes de TALN 2013*.

WEEDS, J. (2003). *Measures and applications of lexical distributional similarity*. Thèse de doctorat, Université du Sussex.